US 20060015942A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0015942 A1**
Judge et al. (43) **Pub. Date: Jan. 19, 2006**

(54) **SYSTEMS AND METHODS FOR CLASSIFICATION OF MESSAGING ENTITIES**

(75) Inventors: **Paul Judge**, Alpharetta, GA (US); **Dmitri Alperovitch**, Atlanta, GA (US); **Matt Moyer**, Lawrenceville, GA (US)

Correspondence Address:
NAGENDRA SETTY
JONES DAY
1420 PEACHTREE ST, NE
SUITE 800
ATLANTA, GA 30309-3053 (US)

(73) Assignee: **CipherTrust, Inc.**

(21) Appl. No.: **11/142,943**

(22) Filed: **Jun. 2, 2005**

Related U.S. Application Data

(63) Continuation-in-part of application No. 10/384,924, filed on Mar. 6, 2003.
Continuation-in-part of application No. 10/373,325, filed on Feb. 24, 2003.
Continuation-in-part of application No. 10/361,091, filed on Feb. 7, 2003.
Continuation-in-part of application No. 10/361,067, filed on Feb. 7, 2003.
Continuation-in-part of application No. 10/094,266, filed on Mar. 8, 2002.
Continuation-in-part of application No. 10/094,211, filed on Mar. 8, 2002.
Continuation-in-part of application No. 10/093,553, filed on Mar. 8, 2002, now Pat. No. 6,941,467.

(60) Provisional application No. 60/625,507, filed on Nov. 5, 2004.

Publication Classification

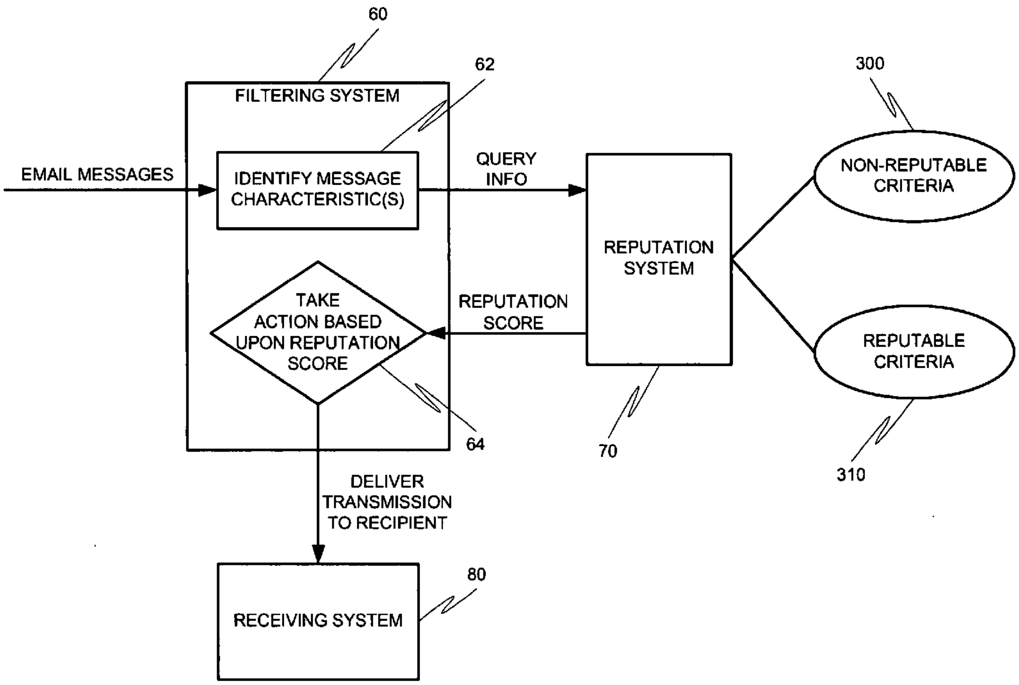
(51) **Int. Cl.**

G06F	*11/00*	(2006.01)
G06F	*11/30*	(2006.01)
G06F	*11/22*	(2006.01)
G06F	*12/14*	(2006.01)
H04L	*9/32*	(2006.01)
G06F	*11/32*	(2006.01)
G06F	*11/34*	(2006.01)
G06F	*11/36*	(2006.01)
G06F	*12/16*	(2006.01)
G06F	*15/18*	(2006.01)
G08B	*23/00*	(2006.01)

(52) **U.S. Cl.** **726/24**; 713/188

(57) **ABSTRACT**

Methods and systems for operation upon one or more data processors for assigning a reputation to a messaging entity. A method can include receiving data that identifies one or more characteristics related to a messaging entity's communication. A reputation score is determined based upon the received identification data. The determined reputation score is indicative of reputation of the messaging entity. The determined reputation score is used in deciding what action is to be taken with respect to a communication associated with the messaging entity.



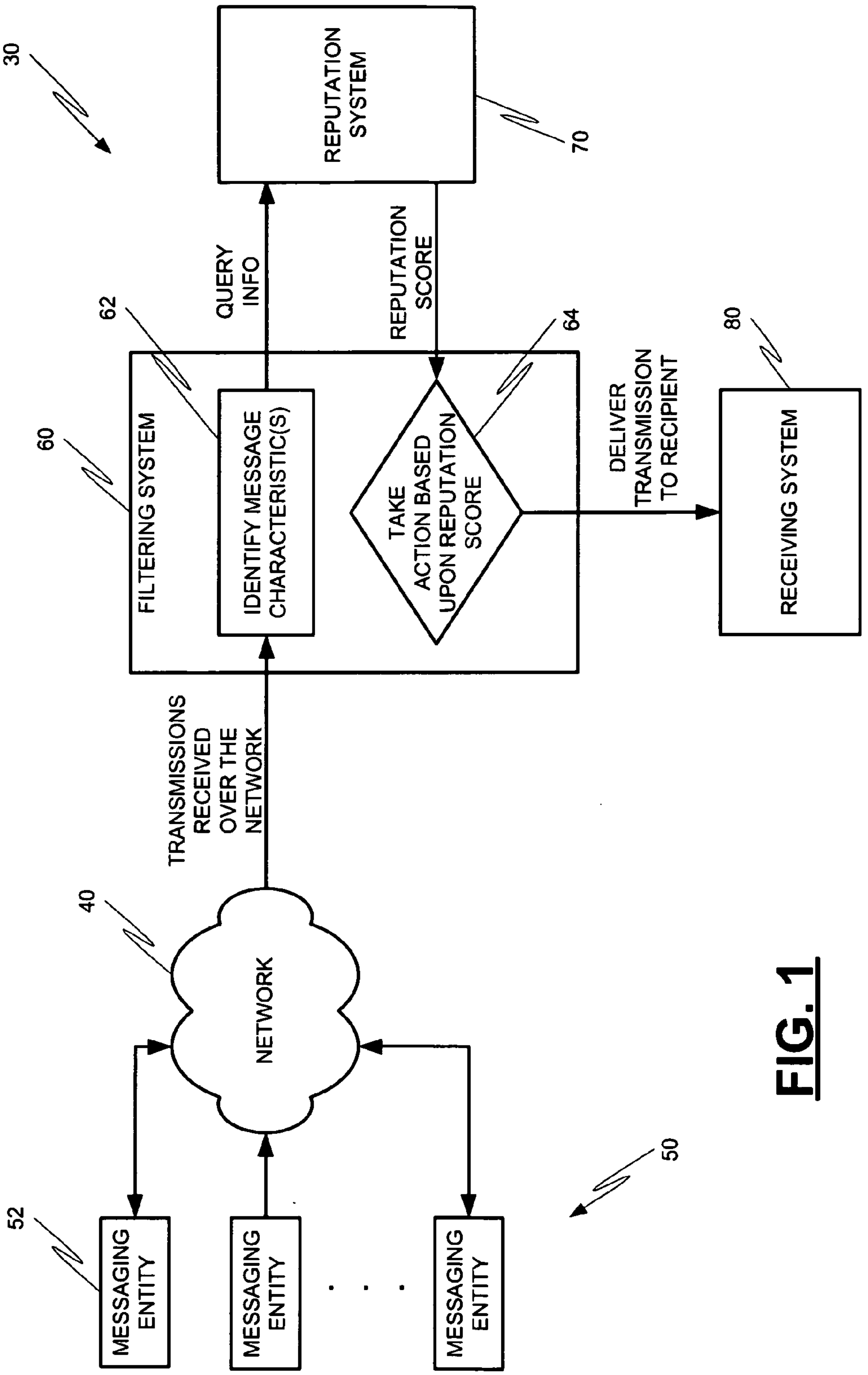

30
MESSAGING ENTITY
MESSAGING ENTITY
. . .
MESSAGING ENTITY
52
50
NETWORK
40
TRANSMISSIONS RECEIVED OVER THE NETWORK
FILTERING SYSTEM
60
IDENTIFY MESSAGE CHARACTERISTIC(S)
62
TAKE ACTION BASED UPON REPUTATION SCORE
64
QUERY INFO
REPUTATION SCORE
REPUTATION SYSTEM
70
DELIVER TRANSMISSION TO RECIPIENT
RECEIVING SYSTEM
80

FIG. 1

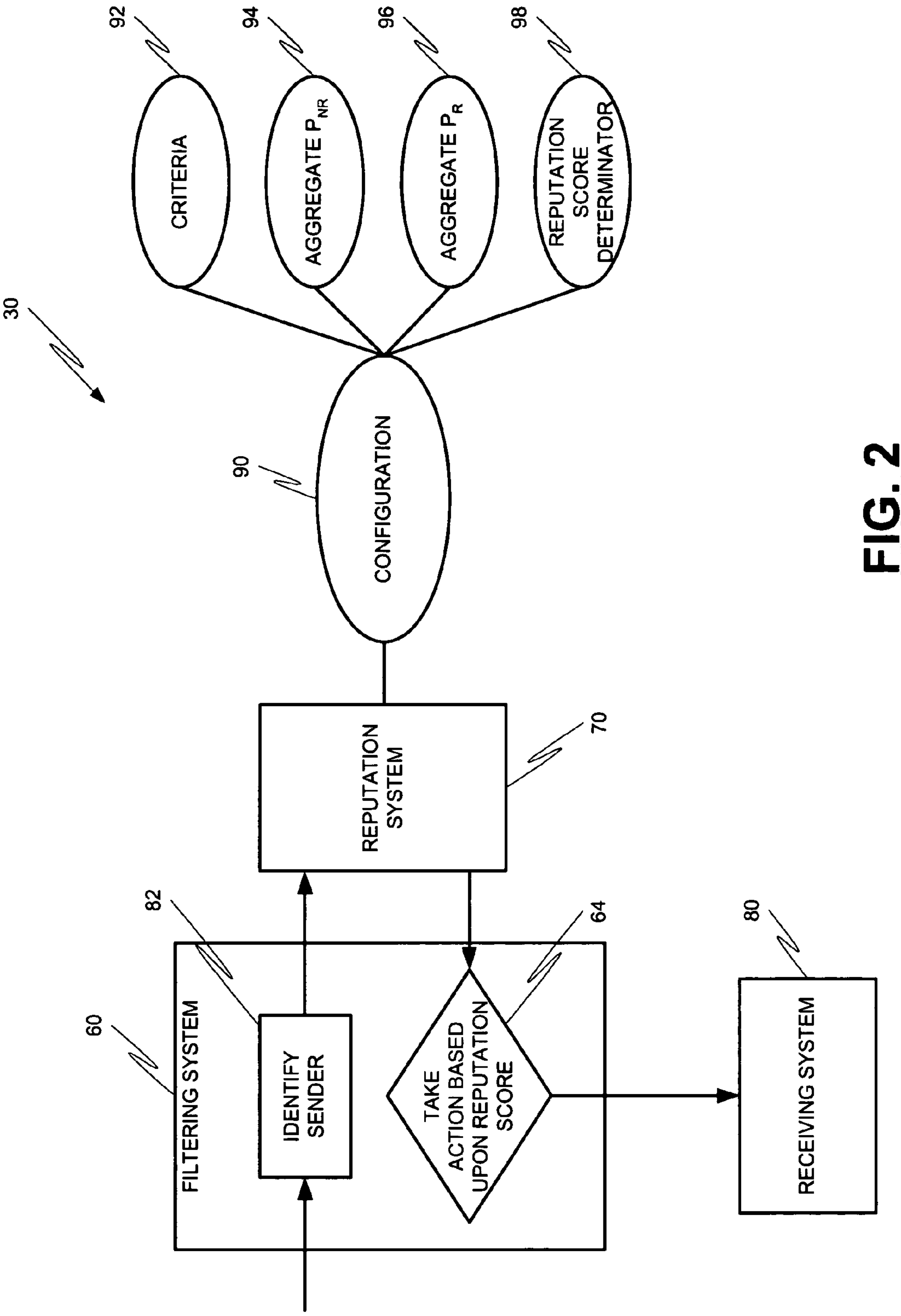
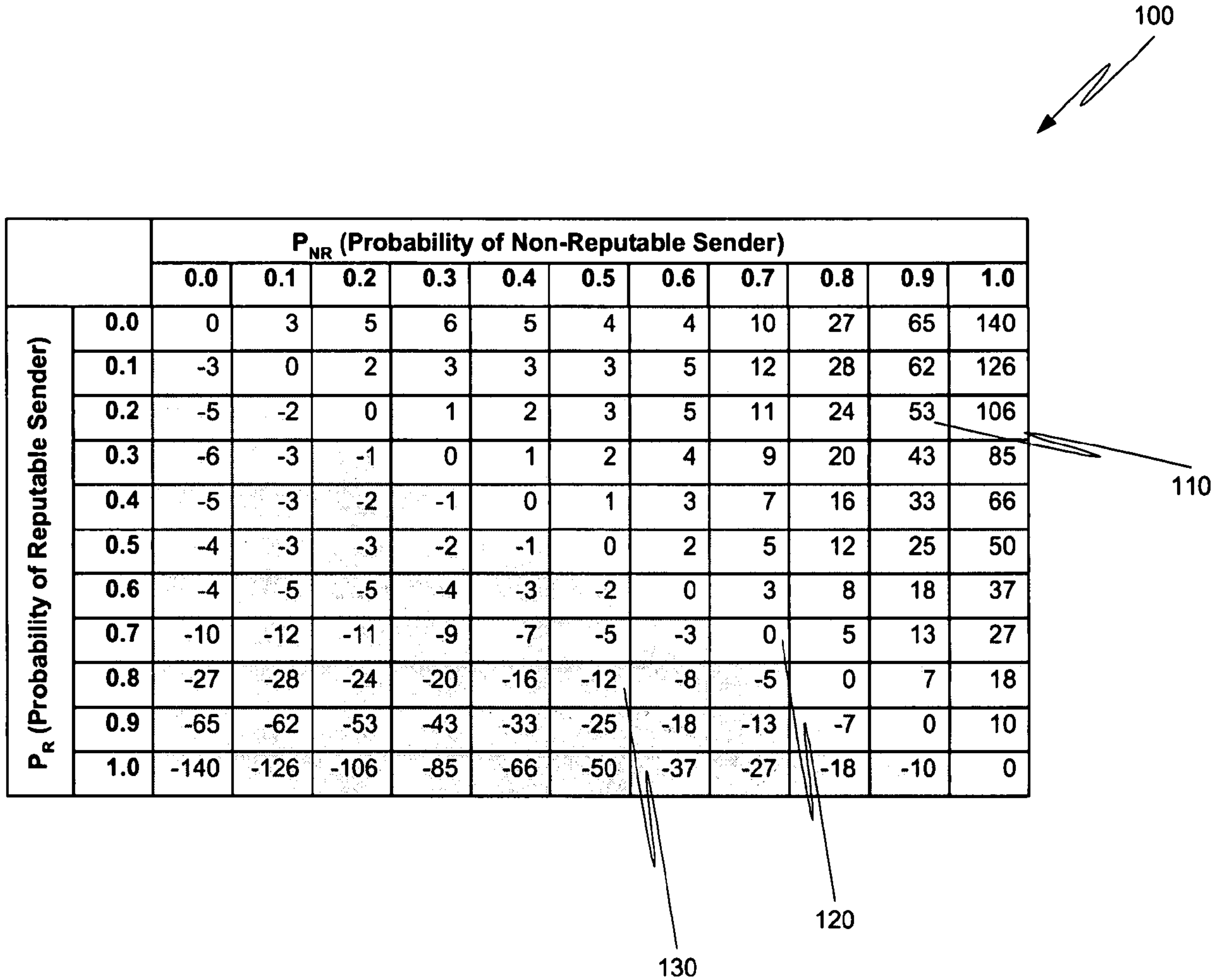

30
92
CRITERIA
94
AGGREGATE P_{NR}
96
AGGREGATE P_{R}
98
REPUTATION SCORE DETERMINATOR
90
CONFIGURATION
REPUTATION SYSTEM
70
60
FILTERING SYSTEM
82
IDENTIFY SENDER
64
TAKE ACTION BASED UPON REPUTATION SCORE
80
RECEIVING SYSTEM

FIG. 2


100
110
120
130

		P_{NR} (Probability of Non-Reputable Sender)										
		0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
P_R (Probability of Reputable Sender)	**0.0**	0	3	5	6	5	4	4	10	27	65	140
	0.1	-3	0	2	3	3	3	5	12	28	62	126
	0.2	-5	-2	0	1	2	3	5	11	24	53	106
	0.3	-6	-3	-1	0	1	2	4	9	20	43	85
	0.4	-5	-3	-2	-1	0	1	3	7	16	33	66
	0.5	-4	-3	-3	-2	-1	0	2	5	12	25	50
	0.6	-4	-5	-5	-4	-3	-2	0	3	8	18	37
	0.7	-10	-12	-11	-9	-7	-5	-3	0	5	13	27
	0.8	-27	-28	-24	-20	-16	-12	-8	-5	0	7	18
	0.9	-65	-62	-53	-43	-33	-25	-18	-13	-7	0	10
	1.0	-140	-126	-106	-85	-66	-50	-37	-27	-18	-10	0

FIG. 3

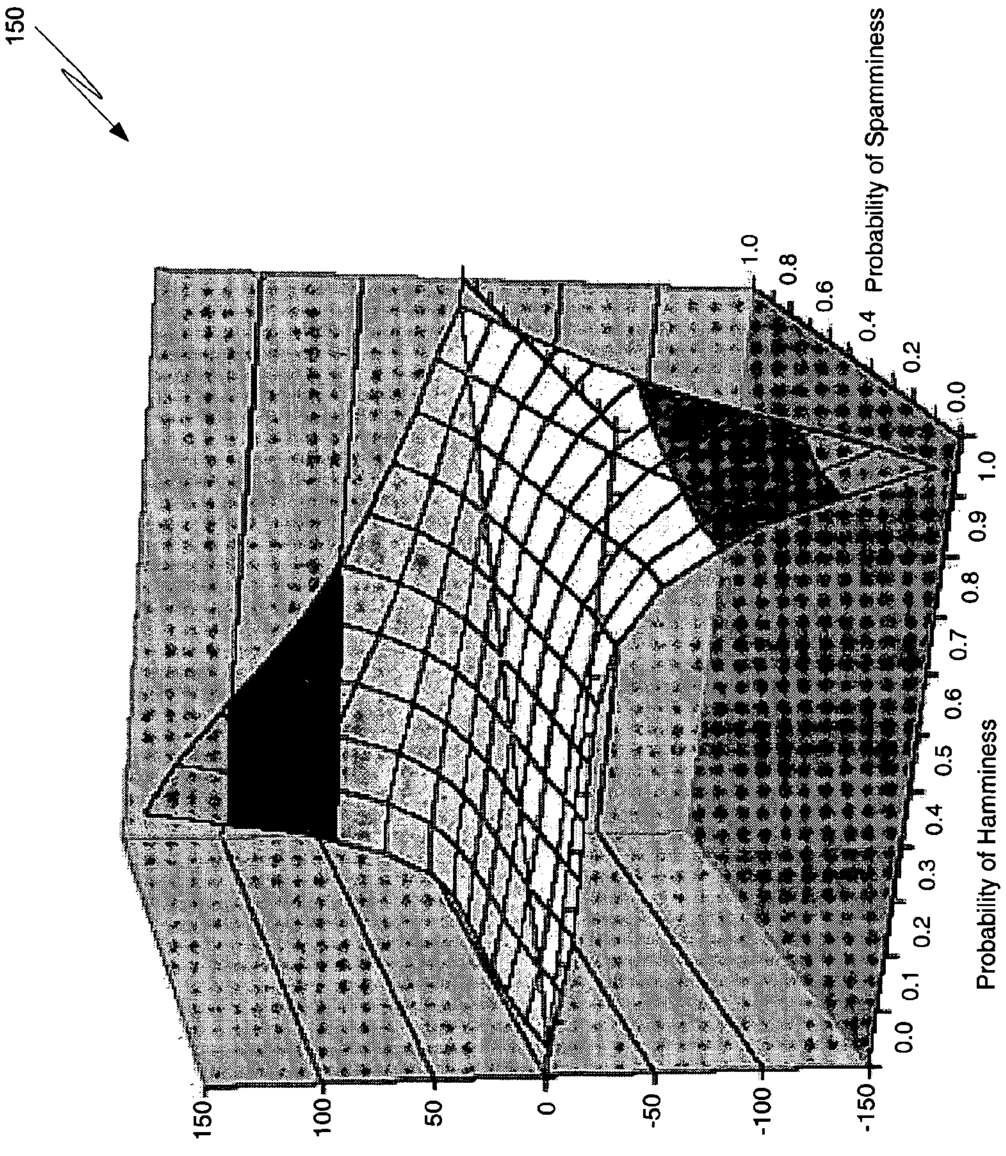

150
150
100
50
0
-50
-100
-150
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Probability of Hamminess
0.0
0.2
0.4
0.6
0.8
1.0
Probability of Spamminess

FIG. 4

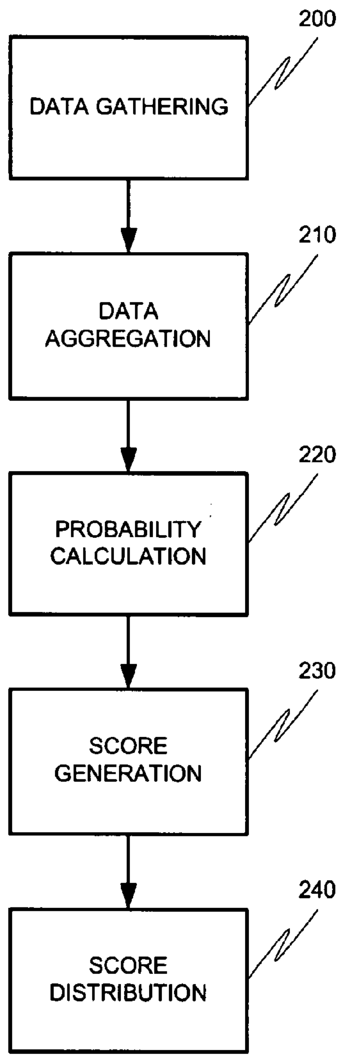

200
DATA GATHERING
210
DATA AGGREGATION
220
PROBABILITY CALCULATION
230
SCORE GENERATION
240
SCORE DISTRIBUTION

FIG. 5

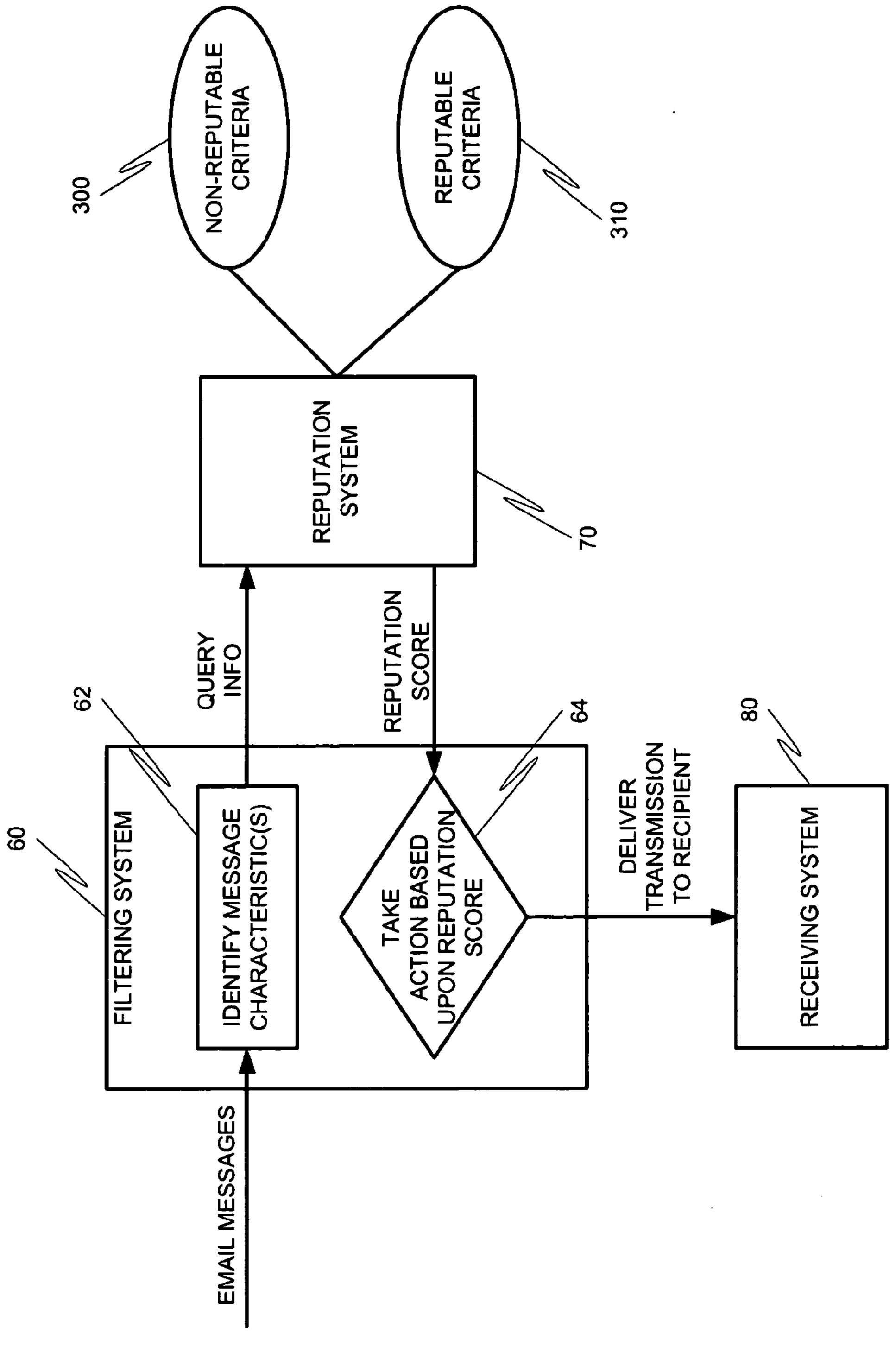

EMAIL MESSAGES
60
FILTERING SYSTEM
62
IDENTIFY MESSAGE CHARACTERISTIC(S)
QUERY INFO
64
TAKE ACTION BASED UPON REPUTATION SCORE
REPUTATION SCORE
REPUTATION SYSTEM
70
300
NON-REPUTABLE CRITERIA
REPUTABLE CRITERIA
310
DELIVER TRANSMISSION TO RECIPIENT
80
RECEIVING SYSTEM

FIG. 6

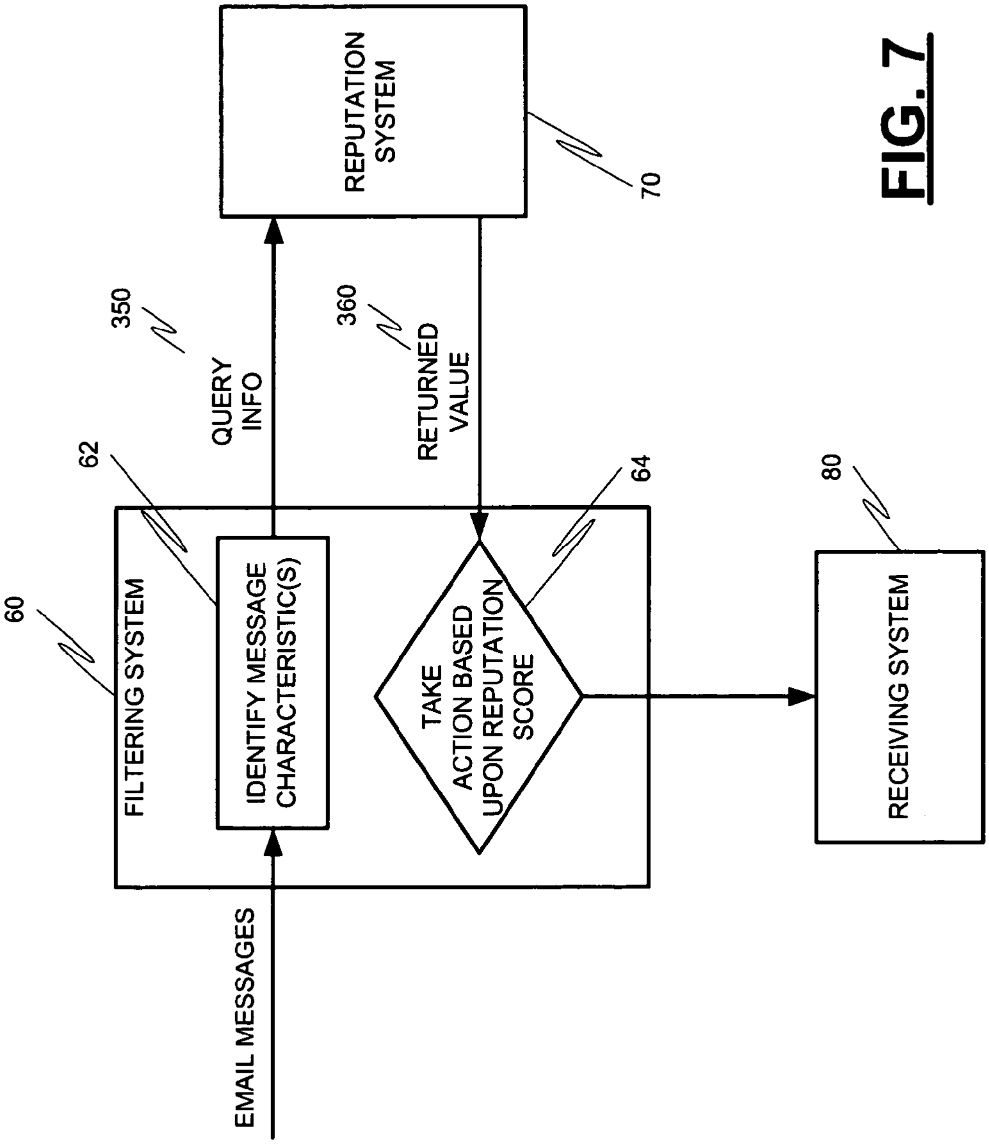

REPUTATION SYSTEM
70
350
QUERY INFO
360
RETURNED VALUE
62
64
80
60
FILTERING SYSTEM
IDENTIFY MESSAGE CHARACTERISTIC(S)
TAKE ACTION BASED UPON REPUTATION SCORE
RECEIVING SYSTEM
EMAIL MESSAGES

FIG. 7

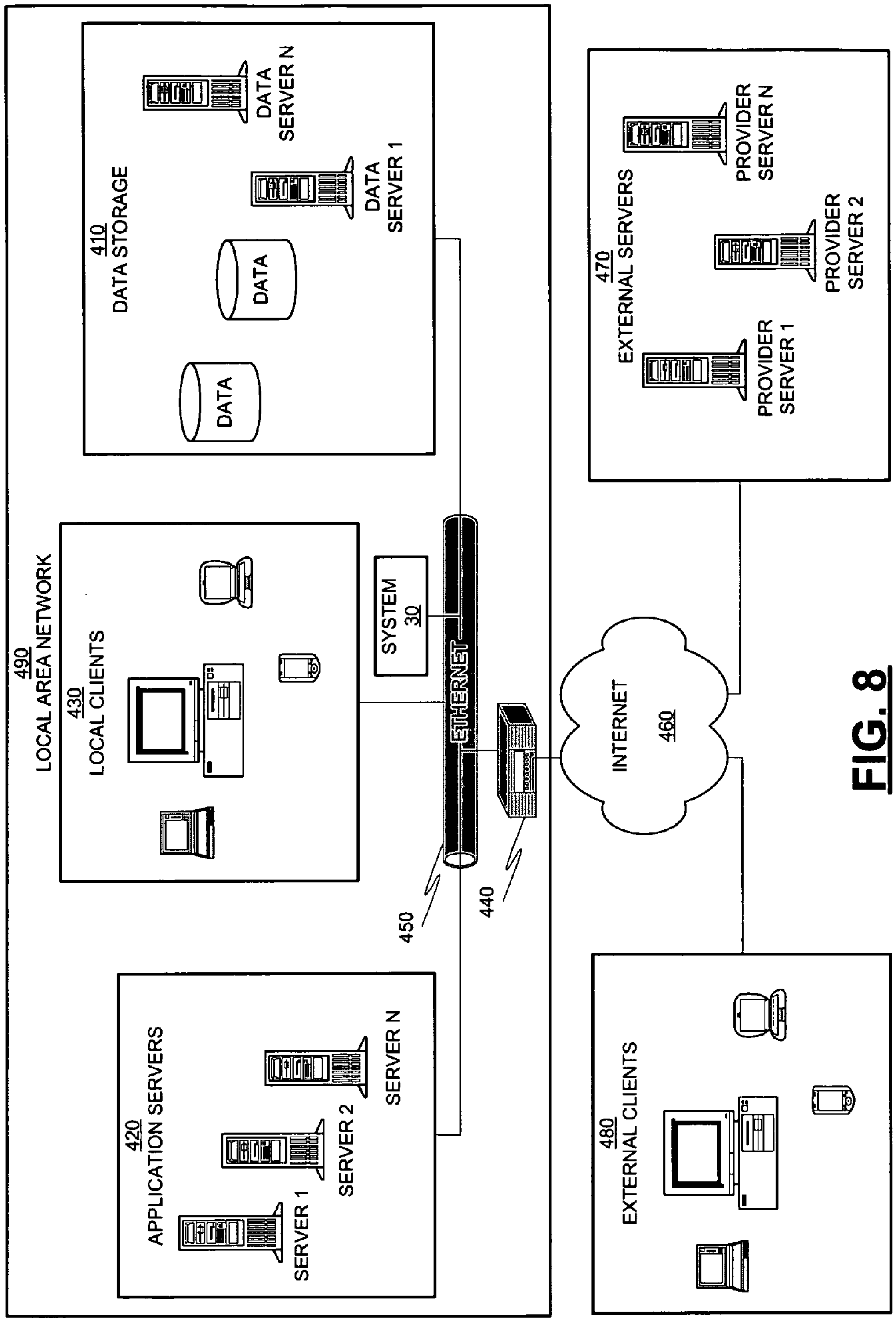

490
LOCAL AREA NETWORK
410
DATA STORAGE
DATA
DATA
DATA SERVER 1
DATA SERVER N
430
LOCAL CLIENTS
SYSTEM 30
ETHERNET
420
APPLICATION SERVERS
SERVER 1
SERVER 2
SERVER N
450
440
INTERNET
460
470
EXTERNAL SERVERS
PROVIDER SERVER 1
PROVIDER SERVER 2
PROVIDER SERVER N
480
EXTERNAL CLIENTS

FIG. 8

SYSTEMS AND METHODS FOR CLASSIFICATION OF MESSAGING ENTITIES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to and the benefit of U.S. Provisional Application Ser. No. 60/625,507 (entitled "Classification of Messaging Entities") filed on Nov. 5, 2004, of which the entire disclosure (including any and all figures) is incorporated herein by reference.

[0002] This application is a continuation-in-part of, and claims priority to and the benefit of, commonly assigned U.S. patent application Ser. No. 10/093,553, entitled "SYSTEMS AND METHODS FOR ADAPTIVE MESSAGE INTERROGATION THROUGH MULTIPLE QUEUES," U.S. patent application Ser. No. 10/094,211, entitled "SYSTEMS AND METHODS FOR ENHANCING ELECTRONIC COMMUNICATION SECURITY," and U.S. patent application Ser. No. 10/094,266, entitled "SYSTEMS AND METHODS FOR ANOMALY DETECTION IN PATTERNS OF MONITORED COMMUNICATIONS," all filed on Mar. 8, 2002, each of which are hereby incorporated by reference in their entirety. This application is also a continuation-in-part of, and claims priority to and the benefit of, commonly assigned U.S. patent application Ser. No. 10/361,091, filed Feb. 7, 2003, entitled "SYSTEMS AND METHODS FOR MESSAGE THREAT MANAGEMENT," U.S. patent application Ser. No. 10/373,325, filed Feb. 24, 2003, entitled "SYSTEMS AND METHODS FOR UPSTREAM THREAT PUSHBACK," U.S. patent application Ser. No. 10/361,067, filed Feb. 7, 2003, entitled "SYSTEMS AND METHODS FOR AUTOMATED WHITELISTING IN MONITORED COMMUNICATIONS," and U.S. patent application Ser. No. 10/384,924, filed Mar. 6, 2003, entitled "SYSTEMS AND METHODS FOR SECURE COMMUNICATION DELIVERY." The entire disclosure of all of these applications is incorporated herein by reference.

BACKGROUND AND SUMMARY

[0003] This document relates generally to systems and methods for processing communications and more particularly to systems and methods for filtering communications.

[0004] In the anti-spam industry, spammers use various creative means for evading detection by spam filters. Accordingly, spam filter designers adopt a strategy of combining various detection techniques in their filters.

[0005] Current tools for message sender analysis include IP blacklists (sometimes called real-time blacklists (RBLs)) and IP whitelists (real-time whitelists (RWLs)). Whitelists and blacklists certainly add value to the spam classification process; however, whitelists and blacklists are inherently limited to providing a binary-type (YES/NO) response to each query. In contrast, a reputation system has the ability to express an opinion of a sender in terms of a scalar number in some defined range. Thus, where blacklists and whitelists are limited to "black and white" responses, a reputation system can express "shades of gray" in its response.

[0006] In accordance with the teachings disclosed herein, methods and systems are provided for operation upon one or more data processors for assigning a reputation to a messaging entity. A method can include receiving data that identifies one or more characteristics related to a messaging entity's communication. A reputation score is determined based upon the received identification data. The determined reputation score is indicative of reputation of the messaging entity. The determined reputation score is used in deciding what action is to be taken with respect to a communication associated with the messaging entity.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] **FIG. 1** is a block diagram depicting a system for handling transmissions received over a network.

[0008] **FIG. 2** is a block diagram depicting a reputation system that has been configured for determining reputation scores.

[0009] **FIG. 3** is a table depicting reputation scores at various calculated probability values.

[0010] **FIG. 4** is a graph depicting reputation scores at various calculated probability values.

[0011] **FIG. 5** is a flowchart depicting an operational scenario for generating reputation scores.

[0012] **FIG. 6** is a block diagram depicting use of non-reputable criteria and reputable criteria for determining reputation scores.

[0013] **FIG. 7** is a block diagram depicting a reputation system configured to respond with a return value that includes the reputation score of a sender.

[0014] **FIG. 8** is a block diagram depicting a server access architecture.

DETAILED DESCRIPTION

[0015] **FIG. 1** depicts at **30** a system for handling transmissions received over a network **40**. The transmissions can be many different types of communications, such as electronic mail (e-mail) messages sent from one or more messaging entities **50**. The system **30** assigns a classification to a messaging entity (e.g., messaging entity **52**), and based upon the classification assigned to the messaging entity, an action is taken with respect to the messaging entity's communication.

[0016] The system **30** uses a filtering system **60** and a reputation system **70** to help process communications from the messaging entities **50**. The filtering system **60** uses the reputation system **70** to help determine what filtering action (if any) should be taken upon the messaging entities' communications. For example, the communication may be determined to be from a reputable source and thus the communication should not be filtered.

[0017] The filtering system **60** identifies at **62** one or more message characteristics associated with a received communication and provides that identification information to the reputation system **70**. The reputation system **70** evaluates the reputation by calculating probabilities that the identified message characteristic(s) exhibit certain qualities. An overall reputation score is determined based upon the calculated probabilities and is provided to the filtering system **60**.

[0018] The filtering system **60** examines at **64** the reputation score in order to determine what action should be taken for the sender's communication (such as whether the communication transmission should be delivered to the communication's designated recipient located within a message receiving system **80**). The filtering system **60** could decide that a communication should be handled differently based in whole or in part upon the reputation scored that was provided by the reputation system **70**. As an illustration, a communication may be determined to be from a non-reputable sender and thus the communication should be handled as Spam (e.g., deleted, quarantined, etc.).

[0019] Reputation systems may be configured in many different ways in order to assist a filtering system. For example, a reputation system **70** can be located externally or internally relative to the filtering system **60** depending upon the situation at hand. As another example, **FIG. 2** depicts a reputation system **70** that has been configured to calculate reputation scores based upon such message characteristic identification information as sender identity as shown at **82**. It should be understood that other message characteristics can be used instead of or in addition to sender identity. Moreover, transmissions may be from many different types of messaging entities, such as a domain name, IP address, phone number, or individual electronic address or username representing an organization, computer, or individual user that transmits electronic messages. For example, generated classifications of reputable and non-reputable can be based upon a tendency for an IP address to send unwanted transmissions or legitimate communication.

[0020] The system's configuration **90** could also, as shown in **FIG. 2**, be established by identifying a set of binary, testable criteria **92** which appear to be strong discriminators between good and bad senders. P $(NR|C_i)$ can be defined as the probability that a sender is non-reputable, given that it conforms to quality/criterion C_i, and P $(R|C_i)$ can be defined as the probability that a sender is reputable, given that it conforms to quality/criterion C_i.

[0021] For each quality/criterion C_i, periodic (e.g., daily, weekly, monthly, etc.) sampling exercises can be performed to recalculate P $(NR|C_i)$. A sampling exercise may include selecting a random sample set S of N senders for which quality/criterion C_i is known to be true. The senders in the sample are then sorted into one of the following sets: reputable (R), non-reputable (NR) or unknown (U). N_R is the number of senders in the sample that are reputable senders, N_{NR} is the number of senders that are non-reputable senders, etc. Then, P $(NR|C_i)$ and P $(R|C_i)$ are estimated using the formulas:

$$P(NR \mid C_i) = \frac{N_{NR}}{N}$$

$$P(R \mid C_i) = \frac{N_R}{N}$$

For this purpose, N=30 was determined to be a large enough sample size to achieve an accurate estimate of P $(NR|C_i)$ and P $(R|C_i)$ for each quality/criterion C_i.

[0022] After calculating P $(NR|C_i)$ and P $(R|C_i)$ for all criteria, the computed probabilities are used to calculate an aggregate non-reputable probability **94**, P_{NR}, and an aggregate reputable sender probability **96**, P_R, for each sender in the reputation space. These probabilities can be calculated using the formulas:

$$P_{NR} = \left(1 - \prod_{i=1}^{N} \begin{cases} 1 - P(NR \mid C_i) & \text{if criterion } i \text{ applies} \\ 1 & \text{otherwise} \end{cases}\right)^{(\text{\# of criteria that apply})}$$

$$P_R = \left(1 - \prod_{i=1}^{N} \begin{cases} 1 - P(R \mid C_i) & \text{if criterion } i \text{ applies} \\ 1 & \text{otherwise} \end{cases}\right)^{(\text{\# of criteria that apply})}$$

In experimentation, the above formulas appeared to behave very well for a wide range of input criteria combinations, and in practice their behavior appears to be similar to the behavior of the formula for correctly computing naïve joint conditional probabilities of "non-reputable" and "reputable" behavior for the input criteria.

[0023] After calculating P_{NR} and P_R for each sender, a reputation score is calculated for that sender using the following reputation function:

$$f(P_{NR}, \quad P_R)=(c_1+c_2P_{NR}+c_2P_R+c_3P_{NR}^2+c_3P^{R2}+ c_4P_{NR}P_R+c_5P_{NR}^3+c_5P_R^3+c_6P_{NR}P_R^2+ c_6P_{NR}^2P_R)((P_{NR}-P_R)^3+c_7(P_{NR}-P_R))$$

[0024] where

[0025] c_1=86.50

[0026] c_2=−193.45

[0027] c_3=−35.19

[0028] c_4=581.09

[0029] c_5=234.81

[0030] c_6=−233.18

[0031] c_7=0.51

It should be understood that different functions can act as a reputation score determinator **98** and can be expressed in many different forms in addition to a functional expression. As an illustration, **FIG. 3** depicts at **100** a tabular form for determining reputation scores. The table shows reputation scores produced by the above function, based on values of P_{NR} and P_R as they each vary between 0.0 and 1.0. For example as shown at **110**, a reputation score of 53 is obtained for the combination of P_{NR}=0.9 and P_R=0.2. This reputation score is a relatively high indicator that the sender should not be considered reputable. A reputation score of 0 is obtained if P_{NR} and P_R are the same (e.g., the reputation score is 0 if P_{NR}=0.7 and P_R=0.7 as shown at **120**). A reputation score can have a negative value to indicate that a sender is relatively reputable as determined when P_R is greater than P_{NR}. For example, if P_{NR}=0.5 and P_R=0.8 as shown at **130**, then the reputation score is −12.

[0032] Reputation scores can be shown graphically as depicted in **FIG. 4** at **150**. Graph **150** was produced by the above function, based on values of P_{NR} and P_R. **FIG. 4** illustrates reputation score determinations in the context of Spam in that the terms P_{NR} and P_R are used respectively as probability of hamminess and probability of spamminess as the probabilities each vary between 0.0 and 1.0.

[0033] As shown in these examples, reputation scores can be numeric reputations that are assigned to messaging entities based on characteristics of a communication (e.g., messaging entity characteristic(s)) and/or a messaging entity's behavior. Numeric reputations can fluctuate between a continuous spectrum of reputable and non-reputable classifications. However, reputations may be non-numeric, such as by having textual, or multiple level textual categories.

[0034] **FIG. 5** depicts an operational scenario wherein a reputation system is used by a filtering system to generate reputation scores. In this operational scenario, a reputation score is computed for a particular sender (e.g., IP address, domain name, phone number, address, name, etc), from a set of input data. With reference to **FIG. 5**, data is gathered at step **200** that is needed to calculate non-reputable and reputable probabilities for a sender. The data is then aggregated at step **210** and used in probability calculations at step **220**. This includes determining, for a sender, non-reputable probabilities and reputable probabilities for various selected criteria. An aggregate non-reputable probability and an aggregate reputable probability are then calculated for each sender.

[0035] After calculating an aggregate non-reputable probability and an aggregate reputable probability for each sender, a reputation score is calculated at **230** for that sender using a reputation function. At step **240**, the sender's reputation score is distributed locally and/or to one or more systems to evaluate a communication associated with the sender. As an illustration, reputation scores can be distributed to a filtering system. With the reputation score, the filtering system can choose to take an action on the transmission based on the range the sender reputation score falls into. For unreputable senders, a filtering system can choose to drop the transmission (e.g., silently), save it in a quarantine area, or flag the transmission as suspicious. In addition, a filter system can choose to apply such actions to all future transmissions from this sender for a specified period of time, without requiring new lookup queries to be made to the reputation system. For reputable senders, a filtering system can similarly apply actions to the transmissions to allow them to bypass all or certain filtering techniques that cause significant processing, network, or storage overhead for the filtering system.

[0036] It should be understood that similar to the other processing flows described herein, the processing and the order of the processing may be altered, modified and/or augmented and still achieve the desired outcome. For example, an optional addition to the step of extracting unique identifying information about the sender of the transmission would be to use sender authentication techniques to authenticate certain parts of the transmission, such as the purported sending domain name in the header of the message, to unforgeable information about the sender, such as the IP address the transmission originated from. This process can allow the filtering system to perform lookups on the reputation system by querying for information that can potentially be forged, had it not been authenticated, such as a domain name or email address. If such domain or address has a positive reputation, the transmission can be delivered directly to the recipient system bypassing all or some filtering techniques. If it has a negative reputation, the filtering system can choose to drop the transmission, save it in a quarantine area, or flag it as suspicious.

[0037] Many different types of sender authentication techniques can be used, such as the Sender Policy Framework (SPF) technique. SPF is a protocol by which domain owners publish DNS records that indicate which IP addresses are allowed to send mail on behalf of a given domain. As other non-limiting examples, SenderID or DomainKeys can be used as sender authentication techniques.

[0038] As another example, many different types of criteria may be used in processing a sender's communication. **FIG. 6** depicts the use of non-reputable criteria **300** and reputable criteria **310** for use in determining reputation scores.

[0039] The non-reputable criteria **300** and reputable criteria **310** help to distinguish non-reputable senders and reputable senders. A set of criteria can change often without significantly affecting the reputation scores produced using this scoring technique. As an illustration within the context of SPAM identification, the following is a list of spamminess criteria that could be used in the reputation scoring of a message sender. The list is not intended to be exhaustive, and can be adapted to include other criteria or remove criteria based upon observed behavior.

[0040] 1. Mean Spam Score: A sender is declared "non-reputable" if a mean spam profiler score of transmissions that it sends exceeds some threshold, W.

[0041] 2. RDNS Lookup Failure: A sender is declared "non-reputable" if reverse domain name system (RDNS) queries for its IP addresses fail.

[0042] 3. RBL Membership: A sender is declared "non-reputable" if it is included in a real-time blackhole list (RBL). (Note: multiple RBLs may be used. Each RBL can constitute a separate testing criterion.)

[0043] 4. Mail Volume: A sender is declared "non-reputable" if its average (mean or median) transmission volume exceeds a threshold, X, where X is measured in transmissions over a period of time (such as, e.g., a day, week, or month). (Note: multiple average volumes over multiple time periods may be used, and each average volume can constitute a separate testing criterion.)

[0044] 5. Mail Burstiness/Sending History: A sender is declared "non-reputable" if its average (mean or median) transmission traffic pattern burstiness (defined by the number of active sending sub-periods within a larger time period, e.g., number of active sending hours in a day or number of active sending days in a month) is less than some threshold, Y, where Y is measured in sub-periods per period. (Note: multiple average burstiness measures over multiple time periods may be used, and each average burstiness measure can constitute a separate testing criterion.)

[0045] 6. Mail Breadth: A sender is declared "non-reputable" if its average (mean or median) transmission traffic breadth (as defined by the percentage of systems that receive transmissions from the same sender during a period of time (such as, e.g., a day, week, or month)) exceeds some threshold, Z. (Note: multiple average breadths over multiple time periods may be used, and each average breadth measure can constitute a separate testing criterion.)

[0046] 7. Malware Activity: A sender is declared "non-reputable" if it is known to have delivered one or more malware codes (such as, e.g., viruses, spyware, intrusion code, etc) during a measurement period (e.g., a day, week, or month).

[0047] 8. Type of Address: A sender is declared "non-reputable" if it is known to be dynamically assigned to dial-up or broadband dynamic host control protocol (DHCP) clients by an internet service provider (ISP).

[0048] 9. CIDR Block Spamminess: A sender is declared "non-reputable" if its IP addresses are known to exist within classless inter-domain routing (CIDR) blocks that contain predominantly "non-reputable" IP addresses.

[0049] 10. Human Feedback: A sender is declared "non-reputable" if it is reported to have sent undesirable transmissions by people analyzing the content and other characteristics of those transmissions.

[0050] 11. SpamTrap Feedback: A sender is declared "non-reputable" if it is sending transmissions to accounts that have been declared as spamtraps and as such are not supposed to receive any legitimate transmissions.

[0051] 12. Bounceback Feedback: A sender is declared "non-reputable" if it is sending bounceback transmissions or transmissions to accounts that do not exist on the destination system.

[0052] 13. Legislation/Standards Conformance: A sender is declared "non-reputable" if it is not conforming to laws, regulations, and well-established standards of transmission behavior in the countries of operation of either the sender and/or the recipient of the transmissions.

[0053] 14. Continuity of Operation: A sender is declared "non-reputable" if it has not operated at that sending location longer than some threshold Z.

[0054] 15. Responsiveness to Recipient Demands: A sender is declared "non-reputable" if it is not responding in a reasonable timeframe to legitimate demands of the recipients to terminate their relationship with the sender to not receive any more transmissions from them.

[0055] The following is a list of "reputable" criteria that could be used in determining the "reputability" of a sender. The list is not intended to be exhaustive, and can be adapted to include other criteria or remove criteria based upon observed behavior.

[0056] 1. Mean Spam Score: A sender is declared "reputable" if the mean spam profiler score of transmissions that it sends falls below some threshold, W.

[0057] 2. Human Feedback: A sender is declared "reputable" if it is reported to have sent only legitimate transmissions by people analyzing transmission flows from that sender, in conjunction with the reputation of the organization that owns those sending stations.

[0058] After computing a reputation grade for each sender in the universe of senders, a reputation classification can be made available via a communication protocol that can be interpreted by the queriers that make use of the reputation system (e.g., DNS, HTTP, etc). As shown in **FIG. 7**, when a query **350** is issued for a sender, the reputation system can respond with a return value **360** that includes the reputation score of that sender, as well as any other relevant additional information that can be used by the querier to make the final judgment on the acceptability of the sender's transmission (e.g., age of the reputation score, input data that determined the score, etc).

[0059] An example of a communication protocol that can be used is a domain name system (DNS) server which can respond with a return value in the form of an IP address: 172.x.y.z. The IP address can be encoded using the formula:

$$IP = 172 \cdot \left(\frac{rep - |rep|}{2 \times rep}\right) \cdot (|rep| div 256) \cdot (|rep| \bmod 256)$$

[0060] The reputation of the queried sender can be deciphered from the return value as follows:

$$\text{rep} = (-1)^{2-x} \times (256y + z)$$

[0061] Therefore, when x=0, the returned reputation is a positive number, and when x=1, the returned reputation is a negative number. The absolute value of the reputation is determined by the values of y and z. This encoding scheme enables the server to return via the DNS protocol reputation values within the range [−65535, 65535]. It also leaves seven (7) unused bits, namely the seven high-order bits of x. These bits can be reserved for extensions to the reputation system. (For example, the age of a reputation score may be communicated back to the querier.)

[0062] The systems and methods disclosed herein may be implemented on various types of computer architectures, such as for example on different types of networked environments. As an illustration, **FIG. 8** depicts a server access architecture within which the disclosed systems and methods may be used (e.g., as shown at **30** in **FIG. 8**). The architecture in this example includes a corporation's local network **490** and a variety of computer systems residing within the local network **490**. These systems can include application servers **420** such as Web servers and e-mail servers, user workstations running local clients **430** such as e-mail readers and Web browsers, and data storage devices **410** such as databases and network connected disks. These systems communicate with each other via a local communication network such as Ethernet **450**. Firewall system **440** resides between the local communication network and Internet **460**. Connected to the Internet **460** are a host of external servers **470** and external clients **480**.

[0063] Local clients **430** can access application servers **420** and shared data storage **410** via the local communication network. External clients **480** can access external application servers **470** via the Internet **460**. In instances where a local server **420** or a local client **430** requires access to an external server **470** or where an external client **480** or an external server **470** requires access to a local server **420**, electronic communications in the appropriate protocol for a given application server flow through "always open" ports of firewall system **440**.

[0064] A system **30** as disclosed herein may be located in a hardware device or on one or more servers connected to the local communication network such as Ethernet **480** and logically interposed between the firewall system **440** and the local servers **420** and clients **430**. Application-related electronic communications attempting to enter or leave the local communications network through the firewall system **440** are routed to the system **30**.

[0065] In the example of **FIG. 8**, system **30** could be configured to store and process reputation data about many millions of senders as part of a threat management system. This would allow the threat management system to make better informed decisions about allowing or blocking electronic mail (e-mail).

[0066] System **30** could be used to handle many different types of e-mail and its variety of protocols that are used for e-mail transmission, delivery and processing including SMTP and POP3. These protocols refer, respectively, to standards for communicating e-mail messages between servers and for server-client communication related to e-mail messages. These protocols are defined respectively in particular RFC's (Request for Comments) promulgated by the IETF (Internet Engineering Task Force). The SMTP protocol is defined in RFC 821, and the POP3 protocol is defined in RFC 1939.

[0067] Since the inception of these standards, various needs have evolved in the field of e-mail leading to the development of further standards including enhancements or additional protocols. For instance, various enhancements have evolved to the SMTP standards leading to the evolution of extended SMTP. Examples of extensions may be seen in (1) RFC 1869 that defines a framework for extending the SMTP service by defining a means whereby a server SMTP can inform a client SMTP as to the service extensions it supports and in (2) RFC 1891 that defines an extension to the SMTP service, which allows an SMTP client to specify (a) that delivery status notifications (DSNs) should be generated under certain conditions, (b) whether such notifications should return the contents of the message, and (c) additional information, to be returned with a DSN, that allows the sender to identify both the recipient(s) for which the DSN was issued, and the transaction in which the original message was sent. In addition, the IMAP protocol has evolved as an alternative to POP3 that supports more advanced interactions between e-mail servers and clients. This protocol is described in RFC 2060.

[0068] Other communication mechanisms are also widely used over networks. These communication mechanisms include, but are not limited to, Voice Over IP (VoIP) and Instant Messaging. VoIP is used in IP telephony to provide a set of facilities for managing the delivery of voice information using the Internet Protocol (IP). Instant Messaging is a type of communication involving a client which hooks up to an instant messaging service that delivers communications (e.g., conversations) in realtime.

[0069] As the Internet has become more widely used, it has also created new troubles for users. In particular, the amount of spam received by individual users has increased dramatically in the recent past. Spam, as used in this specification, refers to any communication receipt of which is either unsolicited or not desired by its recipient. A system and method can be configured as disclosed herein to address these types of unsolicited or undesired communications. This can be helpful in that e-mail spamming consumes corporate resources and impacts productivity.

[0070] The systems and methods disclosed herein are presented only by way of example and are not meant to limit the scope of the invention. Other variations of the systems and methods described above will be apparent to those skilled in the art and as such are considered to be within the scope of the invention. For example, using the systems and methods of sender classification described herein, a reputation system can be configured for use in training and tuning of external filtering techniques. Such techniques may include Bayesian, Support Vector Machine (SVM) and other statistical content filtering techniques, as well as signature-based techniques such as distributed bulk message identification and message clustering-type techniques. The training strategies for such techniques can require sets of classified legitimate and unwanted transmissions, which can be provided to the trainer by classifying streams of transmissions based on the reputation scores of their senders. Transmissions from senders classified as un-reputable can be provided to the filtering system trainer as unwanted, and the wanted transmissions can be taken from the stream sent by the legitimate senders.

[0071] As an illustration, methods and systems can be configured to perform tuning and training of filtering systems utilizing reputation scores of senders of transmissions in sets of trainable transmissions. At least one characteristic is identified about transmissions from senders. The identifying of at least one characteristic can include extracting unique identifying information about the transmissions (e.g., information about the senders of the transmissions), or authenticating unique identifying information about the transmissions, or combinations thereof. Queries are sent to a reputation system and scores are received representing reputations of the senders. Transmissions are classified into multiple categories based on a range a sender's reputation score falls into. Transmissions and their classification categories are passed on to a trainer of another filtering system to be used for optimization of the filtering system.

[0072] As another example, methods and systems can be configured to perform filtering of groups of transmissions utilizing reputation scores of senders of transmissions. Multiple transmissions can be grouped together based on content similarities or similarities in transmission sender behavior. At least one characteristic can be identified about each transmission in the groupings. The identifying of at least one characteristic can include extracting unique identifying information about the transmission (e.g., information about the sender of a transmission), or authenticating unique identifying information about the transmission, or combinations thereof. A query can be sent to the reputation system and receive a score representing reputation of each sender. Groups of transmissions can be classified based on the percentage of reputable and non-reputable senders in the group.

[0073] As another example of the wide variations of the disclosed systems and methods, different techniques can be used for computation of joint conditional probabilities. More specifically, different mathematical techniques can be used for computing the aggregate non-reputable sender probability, P_{NR}, and the aggregate reputable sender probability, P_R, for each sender in the reputation space. As an illustration, two techniques are described. Both techniques use $P(NR|C_i)$ and $P(R|C_i)$, the conditional probabilities of non-reputable and reputable behavior, for each testing criterion C_i. The first technique makes the assumption that all testing criteria are independent. The second technique incorporates the assumption that the testing criteria are not independent. Therefore, the second technique is more difficult to carry out, but produces more accurate results.

[0074] 1. Technique for Independent Testing Criteria

[0075] In the independent case, it is assumed that each criterion C_i is independent of all other criteria. The probability that the sender is non-reputable, P_{NR}, is calculated using the following formula:

$$P_{NR} = \frac{\prod P(NR \mid C_i)}{\prod P(NR \mid C_j) + \prod (1 - P(NR \mid C_j))}$$

where j ranges over all criteria that apply to the sender in question. Similarly, the probability that the sender is a reputable sender, P_R, is calculated using the following formula:

$$P_R = \frac{\prod P(R \mid C_j)}{\prod P(R \mid C_j) + \prod (1 - P(R \mid C_j))}$$

where j ranges over all criteria that apply to the sender in question.

[0076] 2. Technique for Non-Independent Testing Criteria

[0077] In the dependent case, it is assumed that each criterion C_i is not independent of all other criteria, so the analysis must take into account "non-linear" interactions between criteria within their joint probability distribution. To find the correct values for P_{NR} and P_R for a given sender, a table is constructed to represent the entire joint probability distribution. Below is a sample table for a joint distribution of four qualities/criteria.

Case	C_1	C_2	C_3	C_4	P_{NR}	P_R
1	N	N	N	N	N/A	N/A
2	N	N	N	Y	$P(NR \mid C_4)$	$P(R \mid C_4)$
3	N	N	Y	N	$P(NR \mid C_3)$	$P(R \mid C_3)$
4	N	N	Y	Y	$P(NR \mid C_3, C_4)$	$P(R \mid C_3, C_4)$
5	N	Y	N	N	$P(NR \mid C_2)$	$P(R \mid C_2)$
6	N	Y	N	Y	$P(NR \mid C_2, C_4)$	$P(R \mid C_2, C_4)$
7	N	Y	Y	N	$P(NR \mid C_2, C_3)$	$P(R \mid C_2, C_3)$
8	N	Y	Y	Y	$P(NR \mid C_2, C_3, C_4)$	$P(R \mid C_2, C_3, C_4)$
9	Y	N	N	N	$P(NR \mid C_1)$	$P(R \mid C_1)$
10	Y	N	N	Y	$P(NR \mid C_1, C_4)$	$P(R \mid C_1, C_4)$
11	Y	N	Y	N	$P(NR \mid C_1, C_3)$	$P(R \mid C_1, C_3)$
12	Y	N	Y	Y	$P(NR \mid C_1, C_3, C_4)$	$P(R \mid C_1, C_3, C_4)$
13	Y	Y	N	N	$P(NR \mid C_1, C_2)$	$P(R \mid C_1, C_2)$
14	Y	Y	N	Y	$P(NR \mid C_1, C_2, C_4)$	$P(R \mid C_1, C_2, C_4)$
15	Y	Y	Y	N	$P(NR \mid C_1, C_2, C_3)$	$P(R \mid C_1, C_2, C_3)$
16	Y	Y	Y	Y	$P(NR \mid C_1, C_2, C_3, C_4)$	$P(R \mid C_1, C_2, C_3, C_4)$

For a distribution of M criteria, there exist (2M-1) distinct cases within the joint probability distribution. Each case constitutes a particular combination of characteristics. The probability that the sender is non-reputable, P_{NR}, is estimated for each case using the following technique. For each one of the (2M-1) cases, a random sample of N senders is gathered that exhibit the combination of characteristics described by that case. (For this purposes, N=30 is a large enough sample). Each sender is sorted into one of the following sets: reputable (R), non-reputable (NR) or unknown (U). NR is the number of sender in the sample that are reputable senders, N_{NR} is the number of senders that are non-reputable senders, etc. Then, P_{NR} and P_R is estimated using the formulas:

$$P_{NR} = \frac{N_{NR}}{N}$$

$$P_R = \frac{N_R}{N}$$

The sampling of the IP addresses is repeated periodically (e.g., daily, weekly, monthly) to update the joint probability distribution.

[0078] It is further noted that the systems and methods disclosed herein may use articles of manufacture having data/digital signals conveyed via networks (e.g., local area network, wide area network, internet, etc.), fiber optic medium, carrier waves, wireless networks, etc. for communication with one or more data processing devices. The data/digital signals can carry any or all of the data disclosed herein that is provided to or from a device.

[0079] Additionally, the methods and systems described herein may be implemented on many different types of processing devices by program code comprising program instructions that are executable by one or more processors. The software program instructions may include source code, object code, machine code, or any other stored data that is operable to cause a processing system to perform methods described herein.

[0080] The systems' and methods' data (e.g., associations, mappings, etc.) may be stored and implemented in one or more different types of computer-implemented ways, such as different types of storage devices and programming constructs (e.g., data stores, RAM, ROM, Flash memory, flat files, databases, programming data structures, programming variables, IF-THEN (or similar type) statement constructs, etc.). It is noted that data structures describe formats for use in organizing and storing data in databases, programs, memory, or other computer-readable media for use by a computer program.

[0081] The systems and methods may be provided on many different types of computer-readable media including computer storage mechanisms (e.g., CD-ROM, diskette, RAM, flash memory, computer's hard drive, etc.) that contain instructions for use in execution by a processor to perform the methods' operations and implement the systems described herein.

[0082] The computer components, software modules, functions and data structures described herein may be connected directly or indirectly to each other in order to allow the flow of data needed for their operations. It is also noted that software instructions or a module can be implemented for example as a subroutine unit of code, or as a software function unit of code, or as an object (as in an objectoriented paradigm), or as an applet, or in a computer script language, or as another type of computer code or firmware. The software components and/or functionality may be located on a single device or distributed across multiple devices depending upon the situation at hand.

[0083] It should be understood that as used in the description herein and throughout the claims that follow, the meaning of "a,""an," and "the" includes plural reference unless the context clearly dictates otherwise. Also, as used in the description herein and throughout the claims that follow, the meaning of "in" includes "in" and "on" unless the context clearly dictates otherwise. Finally, as used in the description herein and throughout the claims that follow, the meanings of "and" and "or" include both the conjunctive and disjunctive and may be used interchangeably unless the context clearly dictates otherwise; the phrase "exclusive or" may be used to indicate situation where only the disjunctive meaning may apply.

1. A method for operation upon one or more data processors to assign a reputation to a messaging entity, comprising:

receiving data that identifies one or more characteristics related to a messaging entity's communication;

determining a reputation score based upon the received identification data;

wherein the determined reputation score is indicative of reputation of the messaging entity;

wherein the determined reputation score is used in deciding what action is to be taken with respect to a communication associated with the messaging entity.

2. The method of claim 1, wherein the determined reputation score is distributed to one or more computer systems for use in filtering transmissions.

3. The method of claim 1, wherein the determined reputation score is locally distributed to a program for use in filtering transmissions.

4. The method of claim 1, wherein reputation scores include numeric, textual or categorical reputations that are assigned to messaging entities based on characteristics of the messaging entities and their behavior; wherein the numeric reputations fluctuate between a continuous spectrum of reputable and non-reputable classifications.

5. The method of claim 1 further comprising:

determining reputation indicative probabilities based upon the received identification data;

wherein a reputation indicative probability indicates reputability of a messaging entity based upon extent to which the identified one or more communication's characteristics exhibit or conform to one or more reputation-related criteria;

wherein determining the reputation score includes determining the reputation score based upon aggregation of the determined probabilities.

6. The method of claim 5, wherein a type of messaging entity to which reputations are assigned is a domain name, IP address, phone number, or individual electronic address or username representing an organization, computer, or individual user that transmits electronic messages.

7. The method of claim 1 further comprising:

identifying a set of criteria for use in discriminating between reputable and non-reputable classifications;

wherein the criteria include non-reputable criteria and reputable criteria;

using statistical sampling to estimate a conditional probability that a messaging entity displays each criteria;

computing a reputation for each messaging entity, wherein the computing step comprises:

calculating probability that a messaging entity deserves a reputable reputation by computing an estimate of joint conditional probability that the messaging entity is reputable, given the set of criteria that the messaging entity exhibits or conforms to and the individual conditional probability that the messaging entity exhibits or conforms to each such criteria is actually a reputable messaging entity;

calculating the probability that the messaging entity deserves a negative reputation by computing an estimate of joint conditional probability that the messaging entity is non-reputable, given the set of criteria that the messaging entity exhibits or conforms to and the individual conditional probability that the messaging entity exhibits or conforms to each such criteria is actually a non-reputable messaging entity;

computing a reputation for a messaging entity by applying a function to the probabilities.

8. The method of claim 7, wherein the reputation of each messaging entity is encoded within the form of a 32-bit, dotted decimal IP address; said method further comprising:

creating a domain name server (DNS) zone comprising the reputations of all messaging entities in a universe of messaging entities; and

distributing reputations of messaging entities, via the DNS protocol, to one or more computer systems that make use of the reputations for their work.

9. The method of claim 7, wherein the set of criteria are metrics selected from the group: a mean Spam Profiler score; a reverse domain name server lookup failure; membership on one or more real-time blacklists (RBLs); mail volume; mail burstiness; mail breadth; a geographic location; malware activity; a type of address; a classless inter-domain routing (CIDR) block comprising a number of internet protocol addresses identified to send spam; rate of user complaints; rate of honeypot detections; rate of undeliverable transmissions, identified conformance with laws, regulations, and well-established standards of transmission behavior; continuity of operation; responsiveness to recipient demands; and combinations thereof.

10. The method of claim 7, wherein a technique used to compute the joint conditional probabilities is based on probabilistic independence between all criteria.

11. The method of claim 7, wherein a technique used to compute the joint conditional probabilities is based on a joint probability estimation technique.

12. The method of claim 7, wherein a technique used to compute joint conditional probabilities is based on probabilistic non-independence between all criteria.

13. The method of claim 7, wherein the function used to encode the messaging entity reputation within a 32-bit dotted decimal IP address is:

$$IP = 172 \cdot \left(\frac{rep - |rep|}{2 \times rep}\right) \cdot (|rep| div 256) \cdot (|rep| \bmod 256).$$

14. The method of claim 7, wherein classifications of reputable and non-reputable are related to a tendency for an IP address to send unwanted transmissions or legitimate communication.

15. The method of claim 1 further comprising:

determining reputation indicative probabilities based upon the received identification data;

wherein a reputation indicative probability indicates reputability of a messaging entity based upon extent to which the identified one or more communication's characteristics exhibit or conform to one or more reputation-related criteria;

wherein determining the reputation score includes determining the reputation score based upon aggregation of the determined probabilities.

wherein the reputation score is determined based upon applying the aggregation of the determined probabilities to a function;

wherein the function is a function of each of the probabilities that the messaging entity exhibits a reputation-related criterion.

16. A method of performing transmission filtering utilizing reputation scores of transmission sender, the method comprising:

identifying at least one characteristic about a transmission from a sender;

performing a real-time query to the reputation system that includes the transmission characteristic;

receiving a score representing reputation related to the transmission;

performing an action on the transmission from the sender corresponding to the score range of the sender's reputation.

17. The method of claim 16, wherein the action includes at least one of the following actions: rejecting all further transmissions from that sender for a preset period of time or number of transmissions; silently dropping all further transmissions from that sender for a preset period of time or number of transmissions; quarantining all further transmissions from that sender for a preset period of time or number of transmissions; bypassing certain filtering tests for all further transmissions from that sender for a preset period of time or number of transmissions.

18. The method of claim 16, wherein the step of identifying at least one characteristic includes extracting unique identifying information about the transmission, or authenticating unique identifying information about the transmission, or combinations thereof.

19. The method of claim 18, wherein the unique identifying information includes information about the sender of the transmission.

20. A method of performing filtering of groups of transmissions utilizing reputation scores of senders of transmissions, the method comprising:

grouping multiple transmissions together based on content similarities or similarities in transmission sender behavior;

identifying at least one characteristic about each transmission in the groupings;

performing a query to the reputation system and receiving a score representing reputation of each sender;

classifying groups of transmissions based on the percentage of reputable and non-reputable senders in the group.

21. The method of claim 20, wherein the step of identifying at least one characteristic includes extracting unique identifying information about the transmission, or authenticating unique identifying information about the transmission, or combinations thereof.

22. The method of claim 21, wherein the unique identifying information includes information about the sender of a transmission.

23. A method of performing tuning and training of filtering systems utilizing reputation scores of senders of transmissions in sets of trainable transmissions, the method comprising:

identifying at least one characteristic about transmissions from senders;

performing queries to a reputation system and receiving scores representing reputations of the senders;

classifying transmissions into multiple categories based on a range a sender's reputation score falls into;

passing on transmissions and their classification categories to a trainer of another filtering system to be used for optimization of the filtering system.

24. The method of claim 23, wherein the step of identifying at least one characteristic includes extracting unique identifying information about the transmissions, or authenticating unique identifying information about the transmissions, or combinations thereof.

25. The method of claim 24, wherein the unique identifying information includes information about the senders of the transmissions.

26. An article of manufacture comprising a digital signal for transmission using a network; wherein the digital signal includes a query to a reputation process;

wherein the reputation process assigns a reputation to a messaging entity by receiving the query containing data related to a messaging entity's identity;

wherein the identity data is used by the reputation process to determine reputation indicative probabilities;

wherein a reputation indicative probability indicates reputability of a messaging entity based upon extent to which the messaging entity exhibits or conforms to a reputation-related criterion;

wherein a reputation score is determined based upon aggregation of the determined probabilities;

wherein the determined reputation score is indicative of reputation of the messaging entity;

wherein the determined reputation score is used in deciding what action is to be taken with respect to a communication associated with the messaging entity.

27. The digital signal of claim 26, wherein a filtering system generates the digital signal and the reputation process receives the digital signal; wherein the digital signal includes packetized data that is transmitted through the network.

* * * * *